ISSN: 2321-7782 (Online)
Volume 2, Issue 9, September 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

# *Comparative study on Obesity based on ID3 and KNN*

**T.Sivaranjani**
M.Phil Research Schloar
Department of computer science
PSG college of Arts and Science
Coimbatore
Tamil nadu – India

***Abstract: Obesity is base of so many major chronic diseases. In India is rising trend in obesity, progress of urbanization, lifestyle and behavioural changes are cause of Obesity. Lack of awareness about complications of obesity its one of main cause of increasing the prevalence of obesity in India. First consider Overweight or Obese is a disease. This study is proposed to discover the cause of increase obesity and could be predict. This research concentrates upon preventing complications, analysis of causes using a classification data mining technique. The RapidMiner has employed as a software mining tool for analysis. An aim of this research is to find and reduce obesity complications. In this research work proposed two classification algorithms as ID3 and KNN.***

***Keywords: Obesity, prediction, classification, ID3, KNN, RapidMiner.***

## I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years. Data mining can be viewed as a result of the natural evolution of information technology. Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is the process of extracting interesting information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Data mining has been used intensively and broadly by several organizations. The healthcare background is generally supposed as being information more yet knowledge less. There is an affluence of information obtainable within the healthcare systems. However, there is a lack of useful analysis tools to realize hidden relationships and trends in data. Knowledge discovery and data mining have established frequent applications in commerce and scientific domain.

Valuable facts can be exposed from application of data mining techniques in healthcare system. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Obesity is a medical condition in which excess body fat has accumulated to the extent that it may have an adverse effect on health, leading to reduced life expectancy and/or increased health problem. Obesity results from excessive consumption of calories relative to energy expenditure for several years. Approximately 1.2 billion people in the world are overweight and at least 300 million of them are obese. According to the World Health Organization (WHO), obesity is one of the 10 most preventable health risks. Obesity, especially abdominal obesity is a very important risk factor of cardiovascular diseases and some types of cancer. It is also conducive to the development of metabolic and rheumatic diseases, diseases of the liver and billiard ducts, as well as respiratory diseases. Obesity has been thought to simply be related to an imbalance between energy intake and expenditure. However, more recent research has suggested that genetic, physiological, and behavioral factors also play a significant role in the etiology of obesity.

The "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013" report says the population of overweight and obese in globally from 857 million in 1980 to 2.1 billion in 2013. The prevalence of obesity have shown in the below figure 1.

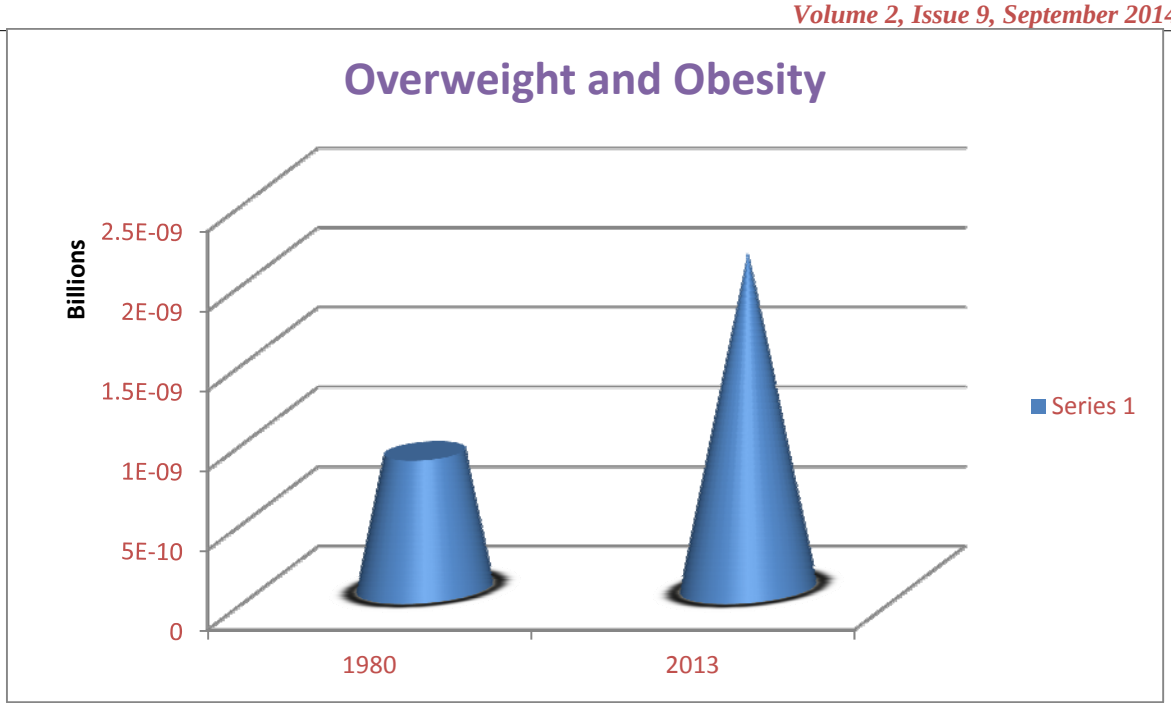



Figure 1 prevalence of overweight and obesity in children and adults during 1980-2013

## II. DATASET DESCRIPTION

- The dataset is of incident instances collected from face-to face using validated questionnaire.
- The *training dataset* consists of 210 records, for this study, the dataset has *nominal, real and integer* data type attributes to be defined.
- The dataset have no missing value to each attribute.
- The dataset has only four classes Normal, Overweight, Obese, and Severely Obese. The classes are sorted based on BMI index value.
- The *training set* is used to apply the classification algorithms and provide the predicted class.

| Sex | BMI | BMI-CLS | Physical Activity | Marital status | Comp-Hrs | Freq-fd intake | Family histroy | Drug-intake | sleeptime |
|---|---|---|---|---|---|---|---|---|---|
| Female | 21.92 | Normal | Little Active | Single | 8 | 4 | Nil | No | 8 |
| Male | 22.85 | Normal | Sedentary | Single | 12 | Nil | Father | No | 7 |
| Female | 26.48 | Overweight | Little Active | Single | 5 | 6 | Father | No | 7 |
| Female | 36.98 | Severely Obese | Sedentary | Single | 4 | 7 | Mother | Yes | 6 |

**2.1 METHODOLOGY DIAGRAM**

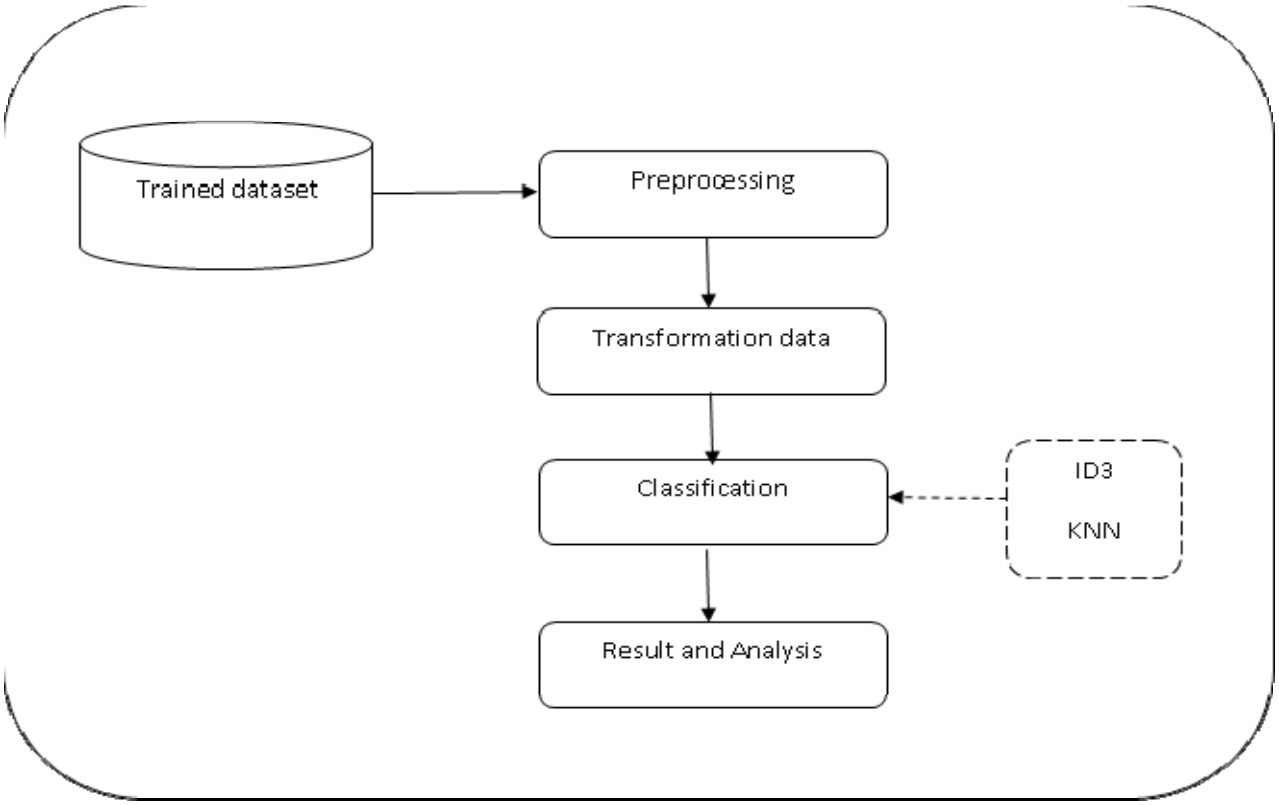



*Figure 2 Methodology diagram of the research work*

In the above figure as describe the each process of this research work. Data pre-processing includes normalization, transformation and selection. The product of data pre-processing is the final training dataset. To predict the causes classification algorithms as ID3and KNN have been applied in the process of classification. In the result and analysis, the performance of classification has to be evaluated and metrics are calculated.

**2.2 CLASSIFICATION**

It is a data mining technique used to map data instances into one of the various predefined categories. It can be used to detect individual attacks but it has high rate of false alarm. Various algorithms like decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques are used for classification techniques. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data.

Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

**2.3 ID3 Algorithm**

ID3 is a simple decision tree erudition algorithm developed by Ross Quinlan (1983) [4]. The basic idea of ID3 algorithm is to create a decision tree of given set, by using top-down greedy search to check each attribute at every tree node.[1].

Given set S, containing these positive and negative targets, the entropy of S related to this Boolean classification is:

Entropy(S) = −P (positive) log2P (positive) − P (negative) log2P (negative)

P (positive): proportion of positive examples in S

P (negative): proportion of negative examples in S

The ID3 algorithm is used to build a decision tree, given a set of non-categorical attributes C1, C2, .., Cn, the categorical attribute C, and a training set T of records.

*Algorithm*

1) Function ID3 (R: a set of non-categorical attributes,

2) C: the categorical attribute,

3) S: a training set) returns a decision tree;

4) begin

5) If S is empty, return a single node with value Failure;

6) If S consists of records all with the same value for the categorical attribute,

7) Return a single node with that value;

8) If R is empty, then return a single node with as value the most frequent of the values of the categorical attribute that are found in records of S; [note that then there will be errors, that is, records that will be improperly classified];

9) Let D be the attribute with largest Gain (D,S)

10) Among attributes in R;

11) Let {*dj*| *j*=1,2, .., *m*} be the values of attribute D;

12) Let {*Sj*| *j*=1,2, .., *m*} be the subsets of S consisting respectively of records with value dj for attribute D;

13) Return a tree with root labeled D and arcs labeled.

14) *d*1, *d*2, .., *dm* going respectively to the trees

15) ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ..,

ID3(R-{D}, C, Sm);

16) end ID3;

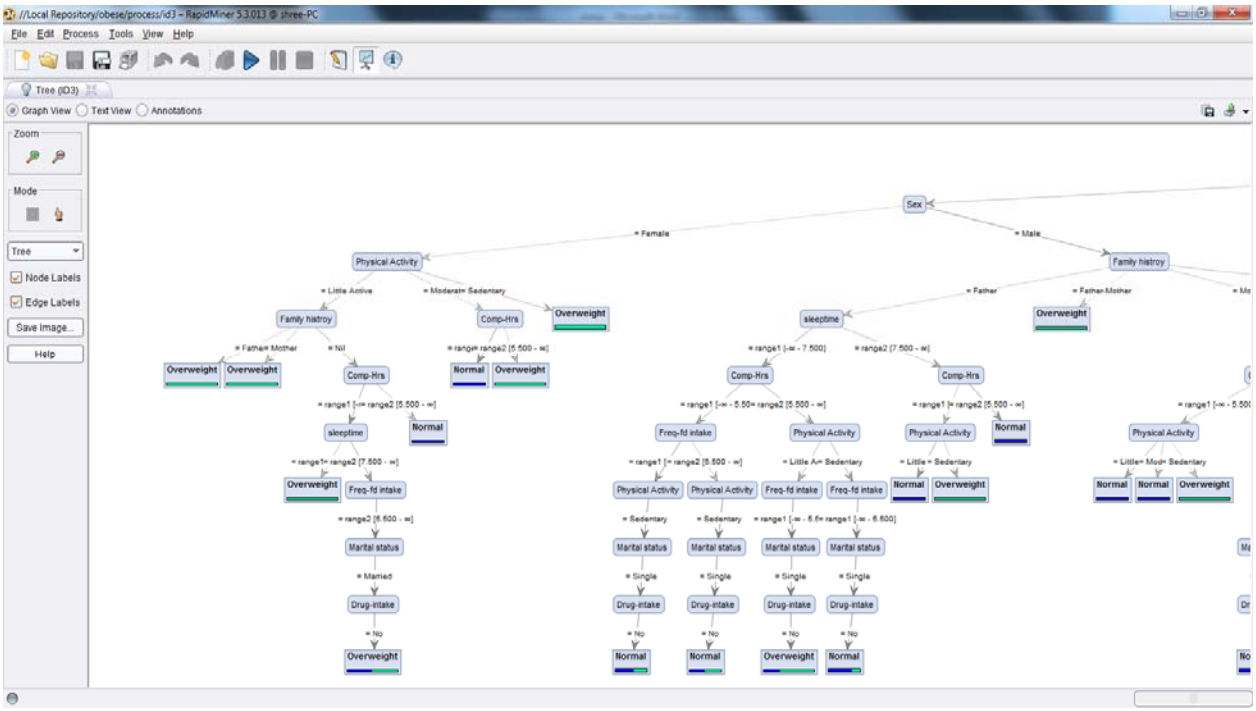



*Fig 3 Classification tree for ID3 algorithm*

Appyling ID3 classification algorithm the classification tree to be constructed based on the constraints prediction classs labeles have classified. The result shown in the fig 3.

**KNN**

K-nearest neighbor ($k$-NN) classifier algorithm is an Instance-based classifier which operates on the premises that classification of unknown instances can be done by relating the unknown to the known according to some distance/similarity function. It is a simple classification algorithm which stores all available cases and classifies the new ones based on a similarity measure i.e. Distance functions. This classification technique has been used already in statistical estimation and pattern recognition using it as a non-parametric technique.

**The KNN Algorithm's pseudo-code:**

**Step 1:** Consider k as the desired number of nearest neighbors,

**Step 2:** S: = p1,...,pn # set of training samples in the form

**Step 3:** p1 = (xi,ci), # $x_i$ is the d-dimensional feature vector of the point $p_i$

$c_i$ is the class that $p_i$ belongs to.

**Step 4:** For each p'=(x',c')

**Step 4.1:** Compute the distance $d(x',x_i)$ between $p'$ and all $p_i$ belonging to S

**Step 4.2:** Sort all points $p_i$ according to the key $d(x',x_i)$

**Step 4.3:** Select the first $k$ points from the sorted list, those are the $k$ closest training samples to $p'$

**Step 4.4:** Assign a class to $p'$ based on majority vote: belonging to S, $I(y = c_i)$

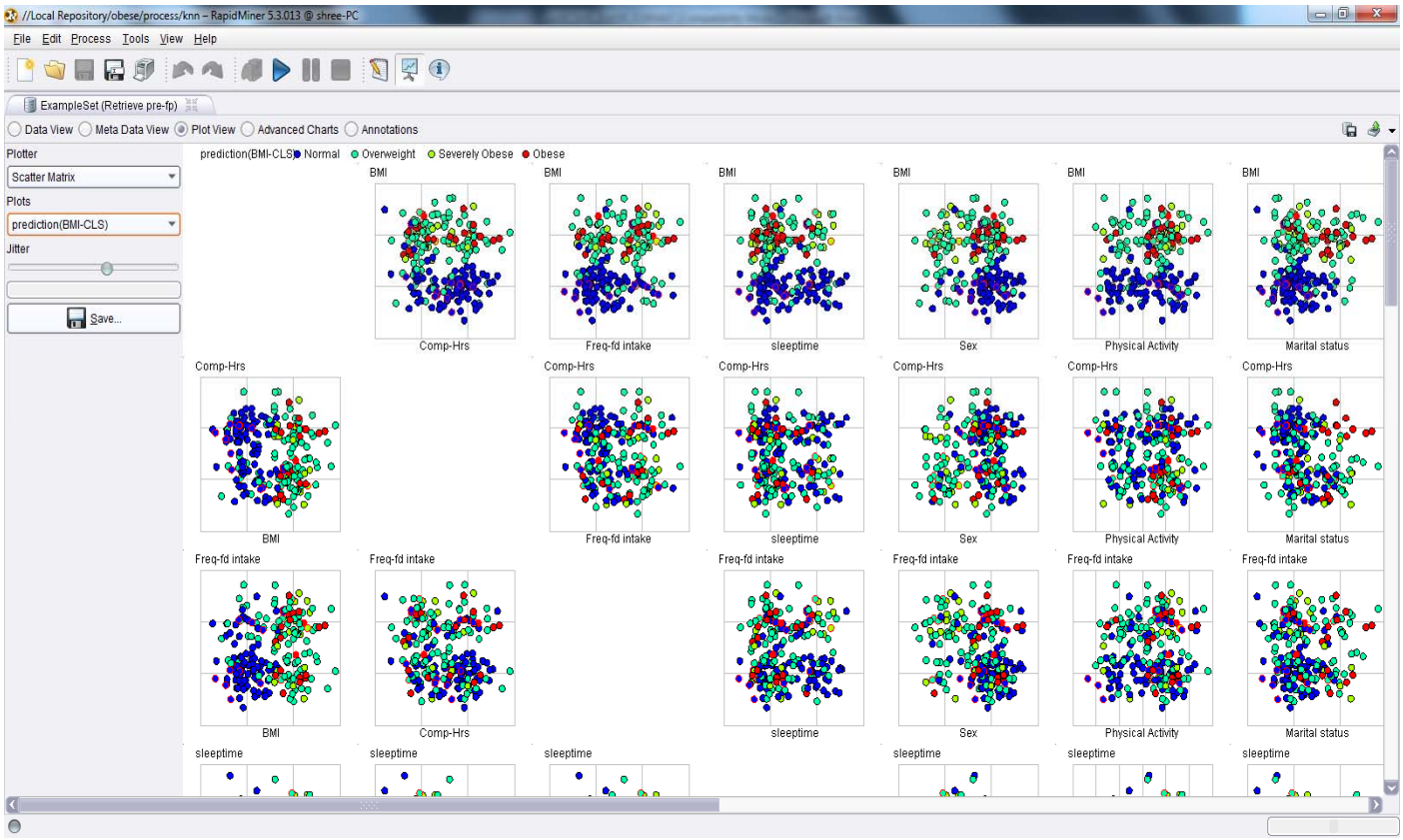



*Fig 4plot viewof Classification algorithm as KNN*

Based on each constraints predicted class labels are classified and it's shown in the figure no 4.Dark blue show normal, green shows over weight, blue shows severely obese and red shows obese.

## III. EVALUATION PERFORMANCE OF CLASSIFICATION

The most popular method for the performance evaluation of classifier are cross-validation, Holdout Method, Random Sub-sampling, k-fold cross validation, Leave one-out Method, Bootstrap, Confusion Matrix, Receiver operating curves (ROC).In this research work, Confusion Matrix is used to evaluate the classification. It is common to call true positive ***hits,*** true negative correct rejections, and false positive *false alarms*, and false negatives *misses*. A number of model performances metric can be derived from the confusion matrix. Perhaps, the most common metric is accuracy defined by the following formula:

$$\textbf{Accuracy} = \frac{TP+TN}{P+N} \qquad \textbf{Error Rate} = \frac{FP+FN}{P+N}$$

Other performance metrics include precision and recall defined as follows:

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad \textbf{Recall} = \frac{TP}{TP+FN} = \frac{TP}{P}$$

| | p' (Predicted) | n' (Predicted) |
|---|---|---|
| P (Actual) | True Positive | False Negative |
| n (Actual) | False Positive | True Negative |

Figure5*: Format of a Confusion Matrix*

To applying Confusion matrix as mathematical model to calculate the performance of classification,. Four metric values to calculated as shown in the below fig 6 and 7 for ID3 and KNN.

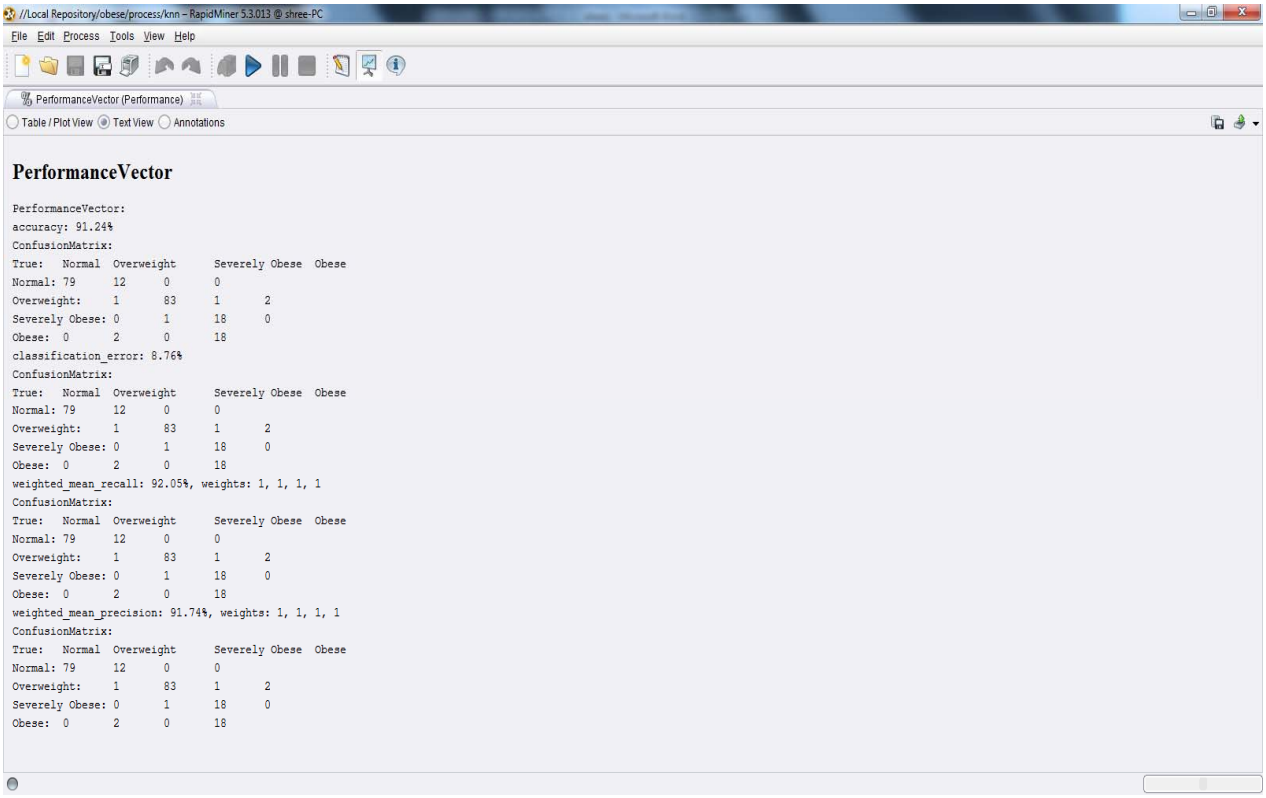



Figure6*:Performance evaluation for the classification algorithm as ID3*

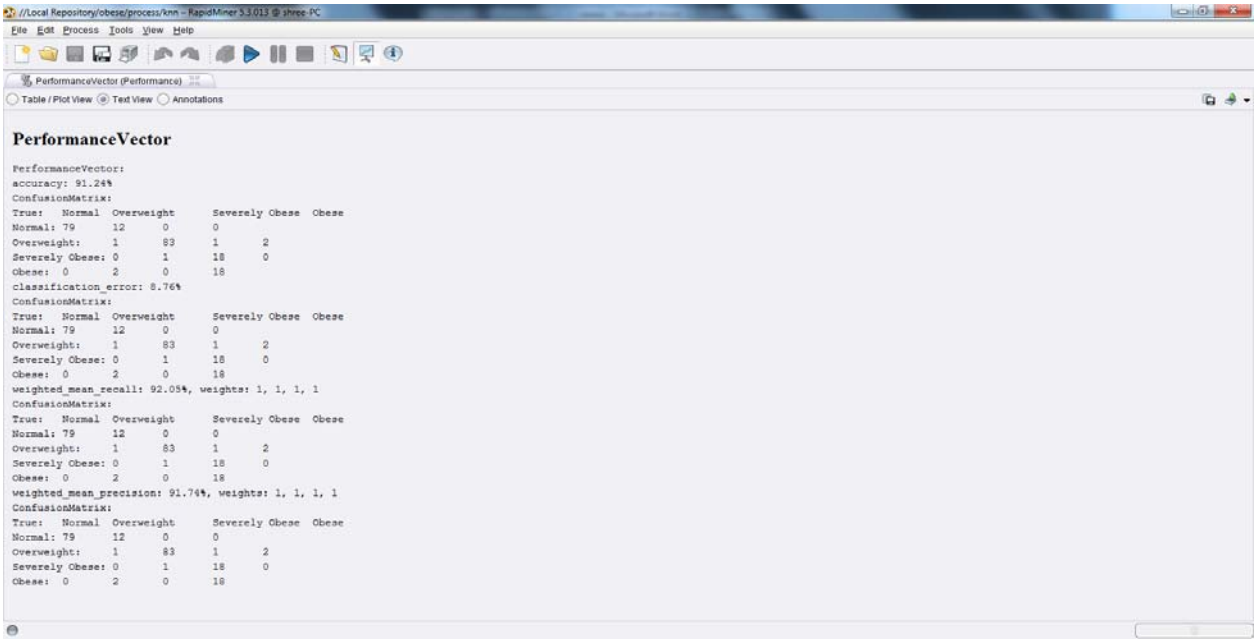
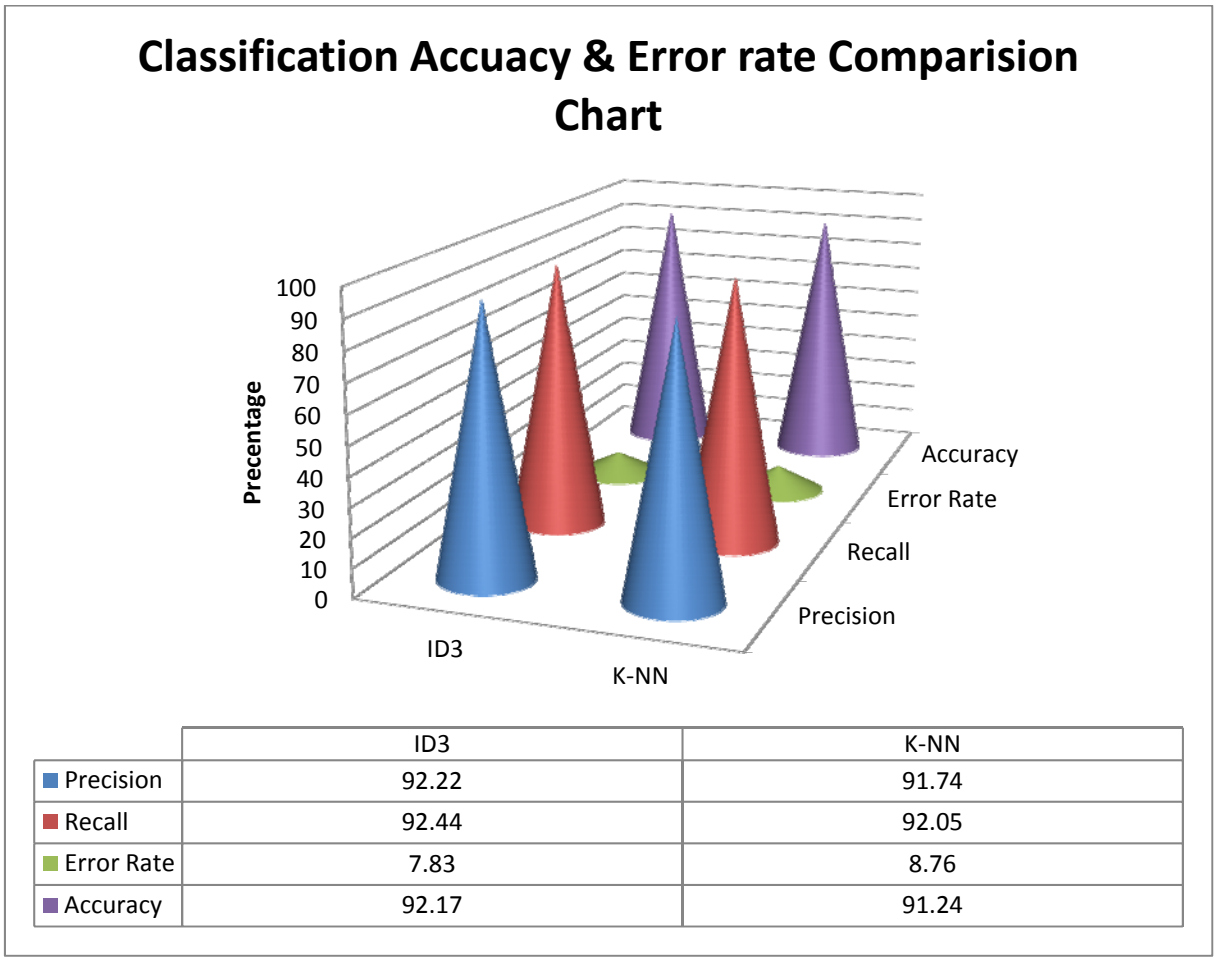



Figure6*: Performance evaluation for the classification algorithm as ID3*

**Comparison chart for the metrics of classification algorithms as ID3 and**




| | ID3 | K-NN |
|---|---|---|
| Precision | 92.22 | 91.74 |
| Recall | 92.44 | 92.05 |
| Error Rate | 7.83 | 8.76 |
| Accuracy | 92.17 | 91.24 |

## IV. CONCLUSION

Obesity is major crisis in globally as well as in India. The major cause of increase the obesity is a sedentary lifestyle. Obesity associated with many other diseases such as heart disease, stroke and diabetic. The prediction of diseases by applying Data Mining tool is a difficult task but it significantly reduces the human efforts and increases the analytical accuracy. Applying an efficient data mining techniques for analysis could reduce the cost and time constraint. The comparison study shows the

interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy like 92.17% for diabetic future risk prediction. On behalf of the results ID3 is provide the best classified data.

## References

1. W. Peng, J. Chen, and H. Zhou, "An Implementation of ID3 – decision Tree Learning Algorithm," University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia.
2. C. F. L. Lima, F. M. de Assis, C. P. de Souza, "Decision Tree based on Shannon, R´enyi and Tsallis Entropies for Intrusion Tolerant Systems," Federal Institute of Maranh˜ao Maracan˜a Campus S˜ao Lu´ıs, MA – Brazil, Federal University of Campina Grande Campina Grande, PB – Brazil, Federal University of Para´ıba Jo˜ao Pessoa, PB – Brazil: Published in The Fifth International Conference on InternetMonitoring and Protection..
3. Arun .K.Pujari "Data Mining Techniques", Universities Press (India) Private Limited,2001.
4. Jiawei Han, Micheline Kamber, "Data mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
5. http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/kNN
6. WHO: World Health Organization. Obesity epidemic puts millions at risk from related diseases.
7. Gortmaker,C., Buzas,J.S., and Sardinha,L. 2003. Social and economic adulthood, New.Eng.J. Med., 359: 1008-1012.
8. James P, Leach R, Kalamara E, Shayeghi M: The Worldwide obesity epidemic. Section I: Obesity, the major health issue of the 21st century. Obes Res 2001; 9:S228-S233.
9. Serdula M, I very D, Coates R, Freedman D, Williamson D, Byers T: Do obese children become obese adults? A review of the literature. Prev Med 1993; 22:167-77.
10. pati NJ, Gaziano TA. Estimating deaths from cardiovascular disease: A review of global methodologies of mortality measurement. Circulation 2013; 127 : 749-56.
11. King H, Aubert RE, Herman W. Global burden of diabetes, 2. 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care 1998; 21 1414-31.

## AUTHOR(S) PROFILE

**T.sivaranjani,** M.phil Research Scholar, Department of Computer Science, PSG college of Arts and Science